David Yang

davidheweiyang@gmail.com | (587)-889-2618 | Calgary, Canada
github.com/davidhy8 | linkedin.com/in/david-h-yang/ | davidhyang.com

SKILLS

Languages & Technologies: Python, R, Java, SQL, Tableau, MS Excel, Git, Minitab, Bash, HTML, CSS, LaTeX, Docker, Spark

Frameworks & Libraries: Data cleaning and Data analysis (Pandas, NumPy, dplyr), Data Visualization (Matplotlib, ggplot2), Machine Learning (scikit-learn, TensorFlow, PyTorch), Data mining (BeautifulSoup), Automation (Selenium), Web development (Flask)

EDUCATION

M.Sc. Mathematics and Statistics – Specialization: Statistics — Sep 2021 – Mar 2024
University of Calgary | GPA 3.7/4.0 | Thesis project: Parallelization of MCMC Phylogenetic Analyses | TA: Calculus
Coursework: Deep Learning, Generalized Linear Models, Statistical Inference, Bayesian Statistics, Theory of Probability

B.Sc. First Class Honours, Cellular, Molecular, and Microbial Biology — Sep 2017 - May 2021
University of Calgary | GPA 3.96/4.00 | Honours project: Eliminating Sampling Bias in SARS-CoV-2 Analysis
Coursework: Computer science I/II, Calculus I/II/III, Linear Algebra I/II, Special Topics in Computer Science, Mathematical Statistics

EXPERIENCE

Data Mining Analyst — Apr 2024 – Present
University of Calgary

- Developed an **ETL pipeline** to retrieve and transform >400,000 scientific papers for Large Language model applications in Python.
- Optimized state-of-the-art **transformer architectures** (BERT) to label sub-chunks of long text and integrated gradient boosting algorithms to overcome BERT's input limitation, achieving >95% recall in identifying scientific papers discussing viral mutations.
- Applied **NER** and **abstractive summarization** with BART to extract and summarize key mutations in flagged scientific papers.

Web Automation Developer – Part-time — Apr 2023 – Present
ADM Lucid Solutions Inc.

- Created automation test scripts with Selenium to validate web application functionality and data integrity (Cucumber, JMeter).

Machine Learning Researcher — Sep 2021 – Mar 2024
University of Calgary

- Pinpointed ~50 out of >30,000 important genomic factors related to Glaucoma disease with R by employing **dimensionality reduction** (regularization, PCA), **data wrangling** (normalization, data imputation), and **statistical testing** techniques (Wald/LRT test, Bootstrapping, Regression) on noisy biological datasets with high dimensionality and multi-collinearity (>30,000 features).
- Generated scientific figures using **data visualization** libraries in R (ggplot2) which elucidated key research findings from **exploratory data analysis** to external institutions leading to the receival of monetary grants valuing greater than $100,000.
- Created an asynchronous parallelization method for the **Markov chain Monte Carlo** (MCMC) algorithm involved in **Bayesian inference** (evolutionary) which reduced computational run-times by more than 2900% (~84 days).
- Identified ~10 key components related to cancer metastasis via **time-series** & **statistical analysis** in R on human blood protein data.

Data Science Researcher — May 2018 – Sep 2021
University of Calgary

- Identified sampling bias in SARS-CoV-2 sequence collection by **analyzing** and **visualizing** COVID-19 data via Python & Tableau**.**
- Devised a novel representative **sampling strategy** based on scientific deductions of COVID-19 and implemented a **data pipeline** involving Python and Perl which reduced sampling bias during SARS-CoV-2 sequence selection (n = >2 million) by around 100%.
- Conducted performance and combability testing for DNA sequencing software that leverages CUDA to parallelize computations.

Chief Information Officer, Co-Founder — Jun 2018 – Aug 2021
Canadian Organization for Undergraduate Health Research

- Leveraged **data analytics** from social media platforms and website traffic to strategically guide recruitment efforts and decision-making, resulting in a 200% increase in hires and a 300% boost in program applicants.

PROJECTS

Electricity forecasting: Time series linear and ridge regression, SARIMA and TimesFM models to predict electricity usage from multivariate time series data with seasonality.

NBA prediction web application: Python pipeline that **web-scrapes** and **preprocesses** >8000 games of NBA data using BeautifulSoup and trains a **neural network** (scikit-learn, TensorFlow**)** to predict NBA win-loss with ~62% accuracy.